Comparison of Empirical Data from Two Honeynets and a Distributed Honeypot Network

Robin Bloomfield, Ilir Gashi, Andrey Povyakalo, Vladimir Stankovic
Centre for Software Reliability, City University London, London, UK
{reb, ec233, andrey, ek274}@csr.city.ac.uk

Abstract

In this paper we present empirical results and speculative analysis based on observations collected over a two month period from studies with two high-interaction honeynets, deployed in a corporate and an SME (Small to Medium Enterprise) environment, and a distributed honeypots deployment. All three networks contain a mixture of Windows and Linux hosts. We detail the architecture of the deployment and results of comparing the observations from the three environments. We analyze in detail the times between attacks on different hosts, operating systems, networks or geographical location. Even though results from honeynet deployments are reported often in the literature, this paper provides novel results analyzing traffic from three different types of networks and some initial exploratory models. This research aims to contribute to endeavours in the wider security research community to build methods, grounded on strong empirical work, for assessment of the robustness of computer-based systems in hostile environments.

1. Introduction

This paper provides details of the research we have conducted with two *honeynets* and a *distributed honeypots* network. One honeynet was deployed in a corporate (City University London) network and the other one in an external network which was purchased from an Internet Service Provider. The latter network (which contains 13 static IP addresses) is meant to simulate an SME (Small to Medium Enterprise) network and therefore allow us to compare the traffic observed in the SME and the corporate environments. Even though honeynet deployment and results are widely reported in the literature we believe that analysis comparing traffic from different environments is scarce and therefore our findings may be of benefit to researchers and practitioners in the field. We have been running honeynets in the corporate network since March 2006 and in the SME network since February 2007. This has enabled us to debug the setups and gain expertise in the deployment and administration of honeynets. Although not reported here, we have implemented a systematic approach to the risk assessment for the networks and forensic procedures. The latter has benefited strongly from our collaboration with contacts in the City of London Police (specifically the High Tech Crime Unit). We classify honeynets according to the attractiveness of the honey and the level of interaction offered. The honeynets reported here are relatively high-interaction honeynets providing potentially attractive *honey* (in the form of computing *resources*) but we have not deployed any *information* honey. We will also report on the attractiveness that the *resources* have on the traffic observed for the corporate network.

Since March 2007 our centre has joined *Leurre.com* [1] distributed honeypots project. Leurre.com make the entire data available to the participating partners (50 partners for the period we analyzed) who are spread around the world. Access to this data enables us to do analysis at a much larger scale than running honeynets on single sites.

In this paper we provide a summary of results observed for all three aforementioned setups (two honeynet networks and the distributed honeypots) in the period 21/May/2007 – 22/July/2007. The main objective of our research was to monitor, study and report the differences in the exposure of the different networks and configurations to malignant traffic, and not to study the "*background radiation*" [2] traffic. We explored the differences in the traffic that is observed in these networks, which are inherently different, rather than imposing various artificial constraints or opening arbitrary ports for the sake of artificially making the honeypots and networks "the same". In the configuration of the hosts in the corporate and SME networks we tried to follow best practices for deploying hosts in corporate or SME environments and we tried to keep the operating systems' configurations

and the applications we deployed in the honeypots as "off-the-shelf" as possible, since we suspect that is how the majority of these installations are done in real deployments. We used the Leurre.com traffic as a reference to see whether or not the malignant traffic to our networks is larger or smaller. We found our traffic is smaller but comparable. In summary, our analysis and modelling has been "exploratory" rather than "explanatory".

The research aims to contribute to the attempt under way in the wider security research community to build rigorous methods for assessing the robustness of a computer-based system in a hostile environment. By modelling system vulnerabilities and counter-measures (both technical and non-technical) we seek to establish credible quantitative *measures* of ease or likelihood of exploitation of these vulnerabilities and of the strength of the counter-measures. Our approach to security assessment is probabilistic. This will allow an integrated view of dependability (where reliability, and most recently safety, have been accepted as requiring a probabilistic approach).

The paper is structured as follows: Section 2 contains an overview of the architecture of the honeynets deployed – full details of the deployment and the various constituent tools of data control and data capture are given in [3]; Section 3 contains a summary of the results observed and some additional *exploratory analysis* of the results in the three networks; Section 4 contains some initial *exploratory models* based on the reported data; Section 5 outlines a brief review of related work and Section 6 discusses the main findings and presents conclusions and provisions for further work.

2. Description of the Test Harness Architecture and Experimental Setup

2.1 The Corporate and the SME Honeynets

The main reference point used for the deployment of the two honeynets (corporate and SME) is the Honeynet Research Alliance "Honeynet Project" [4]. The Honeynet Project develops or incorporates the collection of tools required for data control and data capture: they are provided as a "honeywall", i.e. an installation CD which contains a stripped down version (233 RedHat Packages (*.rpm* files)) of Linux Fedora Core 3 and various security and data collection tools [5]. This is the CD that we used to install and configure the "honeywall" in the two honeynets. We have provided full details of the architecture, the constituent tools and the deployment procedures in [3]. In what follows we will briefly detail the setup to enable the reader to follow the results description and discussions in subsequent sections.

Both the corporate and the SME honeynet have the same basic network and configuration structure. The outline is given in Fig. 1. There are three honeypots (running Fedora Core 6, Windows XP Service Pack 2 and Windows Vista Business Edition respectively). The following applications run on each honeypot:

- Apache web server 2.2.2 [6]
- PostgreSQL database server 8.0.2 [7]
- Open Office 2.2 [8]
- Thunderbird mail client 2 [9]

The applications above were chosen because: they were free and open-source; they were available for each of the operating systems installed on the honeypots. Apart from the TPC-C (Transaction Processing Council - Benchmark C) experimental performance benchmarking database [10], which was deployed in the PostgreSQL server, there is no other content in the honeypots.

The honeywall has three network interface cards – *eth0*, *eth1* and *eth2*; eth0 and eth1 do not have an IP address: they are bridged and the traffic flowing in and out of the honeypots passes through this bridge (therefore the honeywall is not visible to the attackers); eth2 is the management interface: this interface is used by the honeynet administrators to monitor the honeynet; eth2 only responds to requests on ports 443 and 22 (secure HTTP (HTTPS) and Secure Shell (SSH) respectively) and it only responds to a predefined set of trusted hosts.

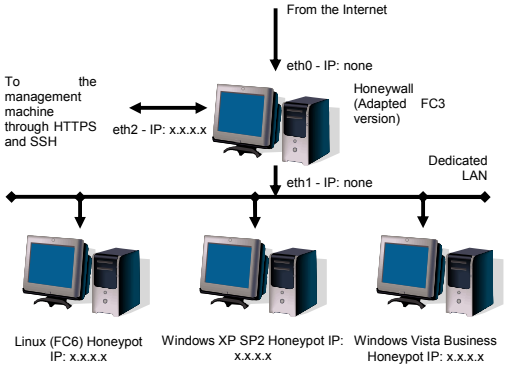

Fig. 1 – The Corporate and the SME honeynet architecture (eth2 not configured in the SME).

The differences in the setup and deployment of the corporate and the SME honeynets are:

- eth2 interface is not configured in the SME honeynet: this is because the ISP package we purchased for running the SME honeynet did not support the assignment of the restrictive features that eth2 should have.
- The corporate honeynet runs behind a firewall (see Fig. 2 for details of the location of the corporate

honeynet in the university network). The firewall is configured to be less restrictive to traffic destined to our honeynet than the rest of the university network. For the SME we could not get any information from the ISP about whether they do any filtering at a higher level before the traffic reaches our segment. Putting the honeypots online without any protection seemed unrealistic. We therefore followed best practice advice directed to home and Small Business users from the UK Get Safe Online site [11], and enabled firewalls on the honeypots, opening only the ports to the applications that were running on the honeypots (80 for Apache, 5432 for PostgreSQL, and 22 for SSH server (in the standard installation of Fedora Core 6 there is an SSH server installed, but not on the Windows honeypots)).

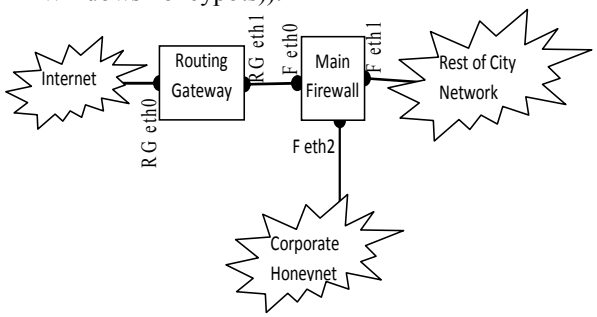

Fig. 2 – The location of the Corporate Honeynet in the City University network.

2.2 Deployment of a Honeynet without Additional Applications in the CORP Network

We have deployed another honeynet in the CORP network with only the bare operating systems deployed in the honeypots (i.e. without applications such as PostgreSQL database server etc., which we installed in the honeypots of CORP and SME honeynets described in the previous section). This was done to allow analysis of the effects that the computing resources (in our case applications) have on the traffic observed. The architecture of the honeynet with only the operating systems in the honeypots is similar to that shown in Fig 1. The only difference is that we do not have a Windows Vista honeypot installed in the honeynet. We will summarise the results of comparison of the two honeynets in the CORP network in Section 3.3.

2.3 The Leurre.com Distributed Honeypots

The Leurre.com project [1] use a different approach to data collection. They use distributed low-interaction honeypots which are dispersed throughout the world. Membership to Leurre.com project is open to everyone as long as the partner organization is willing to sign a non-disclosure agreement, is willing to share the findings with all the other partners and can provide 4 IP addresses. Even though 4 IP addresses are required only one physical host is used. A RedHat operating system is then installed in the host and three other *virtual* hosts are created using *honeyd* [12] and assigned an IP address each. The three hosts emulate Windows NT, Windows 98 and Linux RedHat 7.3 operating systems. The fourth IP address is assigned to the physical host itself. All of the data collected from each honeypot is *flushed* to a centralized data collection database. This database is then available for analysis to all of the participating partners. Since the virtual hosts are created using *honeyd*, the Leurre.com can be described as a *low-interaction* network, i.e. the virtual hosts can be configured to run various services and appear as though they are running various operating systems, but they are not actually real hosts and their capabilities are still limited. However there are advantages from having a worldwide distributed architecture. For example, analysis can be done on how the attacks are distributed at a given time throughout many different locations in the world.

3. Empirical Results

3.1 Summary of Results

In this section we will present a summary of the results observed in the three networks in the period 21/May/2007 until 22/July/2007. For the sake of brevity we use the following abbreviations:

- CORP – Corporate Honeynet
- SME – SME Honeynet
- Leurre – Leurre.com distributed honeypots
- HP - Honeypot
- Avg. - Average

We have also analyzed the traffic from attacking hosts that have been observed in more than one network. Table 1 contains a summary of the results for all three networks. For example, in the last section of Table 1 we can see that a total of 225 attacking hosts were observed in all three networks. This constitutes only 0.09% of the hosts seen in Leurre, but 12.84% of the hosts in CORP.

If we look at the pair-wise comparisons of the networks we can see that a high percentage of attacking IP addresses that were observed in CORP were also observed in the Leurre network (90.09%). This figure is lower for the SME attackers found in Leurre (34.6%). A possible explanation for the higher percentage of commonality of attackers of CORP and Leurre (compared with SME and Leurre) is that most

of the Leurre honeypots are located in large university or research institutions ("corporate" networks); therefore one would expect the attackers of these two networks to be in some way similar.

Full details and a more detailed empirical analysis of the observations are provided in a technical report [13]. The following are some of the more interesting observations to note about these data:

- *Why is there variation in the overall traffic volume in the three networks*? Due to the higher number of hosts that are active in Leurre network (150 in total) the total number of packets observed in Leurre is significantly higher than in CORP or SME. SME network has the least amount of traffic. A possible explanation for the smaller number of packets in the SME network is that larger corporate networks, whose IP addresses are often public (certainly true for CORP honeynet that we run on the City University London network), may be more tempting for attackers (due to, for example, possible access to classified materials on the corporate networks; or, for attackers who are after resources, the possible availability of a larger number of hosts) than SME networks whose identity may not be known (if we try to resolve the identity of the honeypots in our SME network using tools such as *whois*, the owner is shown to be the ISP from whom we purchased the subnet).
- *Which honeypot ports were scanned/attacked most frequently*? Ports 80, 135, 5900 (used for Virtual Network Computing) and Microsoft SQL Server ports (1433 and 1434), as well as ICMP traffic that does not use a port abstraction, feature in the top 10 (ranked by total number of packets exchanged) in all three datasets. Similarly, five ports feature in the top 10 in two networks: 22, 139, 445, 1026 and 2967. These observations suggest that most of the packet exchanges are with ports that have known, but mostly patched, vulnerabilities.
- *From which countries were the attacking IP addresses*? The USA IP addresses exchange most of the packets with any of the three networks. In terms of the number of distinct attacker IPs, China features in the top 5 for any of the three networks. It may be surprising to some that the number of attacking IPs from Russia (and Eastern Europe in general) was relatively small.

Table 1 – Summarized packet and IP counts for all traffic or only traffic from attacker IPs observed in more than one network.

	CORP			SME			Leurre		
	Attacking IPs Count	Packet Count	Avg. per Attac. IP	Attacking IPs Count	Packet Count	Avg. per Attac. IP	Attacking IPs Count	Packet Count	Avg. per Attac. IP
Corp&SME Only	244	*7,787*	***31.91***	244	*9,237*	***37.86***			
Totals for all IPs	1,752	402,134	***229.53***	10,028	86,896	***8.67***			
Ratio of totals	***10.67%***	***1.12%***		***2.82%***	***3.96%***				
Corp&Leurre Only	1,090	362,268	***332.36***				1,090	207,363	***190.24***
Totals for all IPs	1,752	402,134	***229.53***				243,188	29,355,199	***120.71***
Ratio of totals	***62.21%***	***90.09%***					***0.45%***	***0.71%***	
SME&Leurre Only				3,470	34,817	***10.03***	3,470	556,510	***160.38***
Totals for all IPs				10,028	86,896	***8.67***	243,188	29,355,199	***120.71***
Ratio of totals				***34.60%***	***40.07%***		***1.43%***	***1.90%***	
All Three	225	5,501	***24.45***	225	7,582	***33.70***	225	82,356	***366.03***
Totals for all IPs	1,752	402,134	***229.53***	10,028	86,896	***8.67***	243,188	29,355,199	***120.71***
Ratio of totals	***12.84%***	***1.37%***		***2.24%***	***8.73%***		***0.09%***	***0.28%***	

3.2 "Persistent" vs. "Prolific" Attacking IP Addresses

We will now look more closely at the traffic originating from the top 10 most "persistent" attacking IPs (the ones that have exchanged the most packets with the honeypots in the network) and the top 10 most "prolific" attacking IPs (the ones that have attacked the most honeypots). This analysis aims to clarify the understanding about the different strategies that attackers are using. Are they going for "depth" attacks, i.e. concentrating their efforts on a relatively small number of hosts; or for "breadth" attacks, i.e. scanning/attacking as high number of hosts as possible.

3.2.1 Analysis of the "Persistent" IP Addresses. Fig. 3 shows normalised (per honeypot in each network) count of packets exchanged with the three networks by the top 50 most "persistent" hosts. Note that in the figure we have an ordering of the most persistent

attacking hosts in each network. Therefore they are not necessarily the same. The y-axis is drawn in an exponential scale.

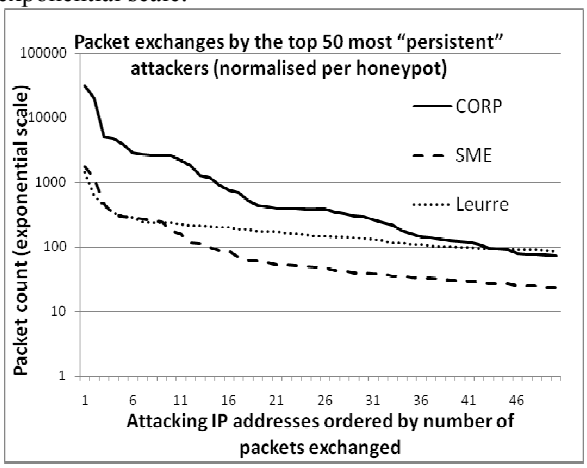

Fig. 3 - Packet exchanges by the top 50 most "persistent" attackers.

From Table 1 we can see that the total number of attacking hosts in each network is much greater than 50 (1,752; 10,028 and 243,188, for CORP, SME and Leurre respectively). Therefore the number of highly persistent hosts in each network is relatively low and, as a result, the tails of the three lines in Fig. 3 are very long.

3.2.2 Analysis of the "Prolific" IP addresses in the Leurre Network. Table 2 shows the top ten attacking IPs[1] in terms of number of Honeypots they attacked. The table only shows the IPs from the Leurre network because significantly higher number of Honeypots exist in this network, making the analysis more interesting. We can see that 7 out of 10 most prolific attacking IP addresses are from China.

Fig. 4 shows the pattern in which the honeypots were discovered by the 50 most prolific IP addresses (i.e. the graph shows the elapsed times, in the observation period, of the first packet exchanges with each honeypot). We can see that the honeypot discovery patterns of some of the attackers have a convex shape (i.e. the attackers discover a lot of honeypots initially in relatively short period of time and then take longer to discover the remaining honeypots), whereas a few have a concave shape (taking longer initially and then increasing the discovery rate). This may be due to:

- the observation period being limited to 63 days and consequently some of the discovery patterns may be cut short.
- different attackers using different attack strategies, e.g. concave: discover new honeypots slowly and concentrate attacks against a few hosts, and then restart discovery/scanning again.
- us seeing only part of the global picture of host discovery patterns by these attackers, i.e. other worldwide hosts (which are not covered by the 150 hosts in the Leurre network) are maybe being discovered by these attackers in between the discovery of the Leurre honeypots.

[1] The attacking hosts' IP addresses are not shown due to the restrictions of Leurre.com non-disclosure agreement.

Table 2 – Top 10 most 'prolific' attacking IP addresses (anonymised) in the Leurre network. The top two hosts are from the same C-class subnet.

Foreign IP	Country	Number of HPs attacked	% of Total (150) HPs	Packet count
A1	China	120	80.00%	3,909
A2	China	113	75.33%	1,394
B	China	111	74.00%	403
C	China	106	70.67%	2,753
D	USA	105	70.00%	2,013
E	N'lands	105	70.00%	887
F	USA	104	69.33%	7,222
G	China	102	68.00%	1,200
H	China	102	68.00%	285
I	China	101	67.33%	3,464

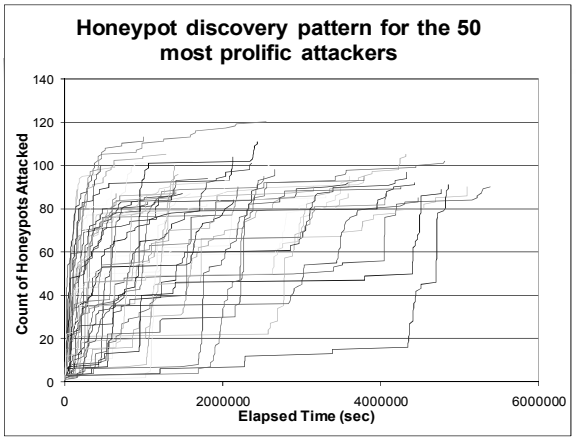

Fig. 4 – Honeypot discovery pattern for the 50 most prolific attackers.

Another interesting observation to note about the prolific attackers is that when they are scanning honeypots that reside in the same network, they seem invariably to do so sequentially in ascending numerical order of the IP addresses per given site (there are exactly 3 hosts per site in Leurre (for the total of 50 sites there are 150 honeypots)). A related study with Leurre network [14] also observed this phenomenon.

From Table 2 we saw that the top two most prolific IP addresses are from the same class C network in China. Fig. 5 shows the pattern in which the honeypots were discovered by these two IP addresses. We can see that the two patterns are quite similar for the discovery of the first 100 honeypots. The time to discover the remaining ones is much longer for both attackers (the "dotted line" one discovers less, due to the observation period closing on the 22nd of July 2007). The "solid line" discovery pattern started on the 25th of May 2007; the "dotted line" discovery pattern started on the 5th of July 2007. It seems highly plausible that the two attacking IP addresses might either belong to the same attacker who obtained different IPs in these two periods, or are two identically compromised/hijacked hosts.

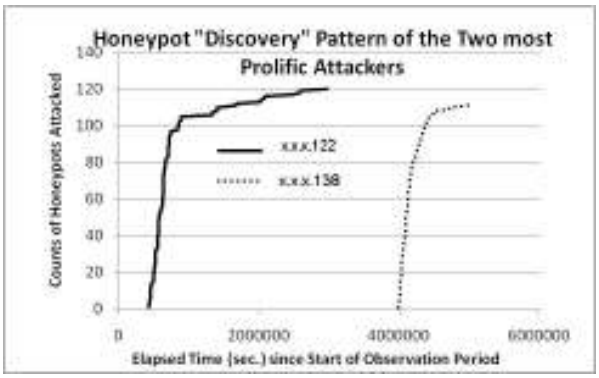

Fig. 5 – The Honeypot discovery pattern for the two most prolific attacking IPs.

3.2.3 Comparison of the "Persistent" and "Prolific" traffic. Table 3 contains a comparison of the traffic received by the top 10 most persistent and the top 10 most prolific attackers. For Leurre these 10 hosts are mutually exclusive, for CORP there are 2 hosts in the top 10 of both persistent and prolific lists whereas for SME this number is 6 hosts. The reason for the higher overlaps for SME and CORP is because the number of honeypots in these two networks is very small.

Table 3 – Top 10 hosts: 'prolific' vs. 'persistent' attacking hosts traffic comparison.

		Packet Count	% of total traffic	Avg. number of HN ports attacked	Avg. number of hosts attacked
CORP	Top 10 Persistent	310,475	77.00	1010.2	*2.5 / 4*
	Top 10 Prolific	172,916	43.00	1010	*4 / 4*
SME	Top 10 Persistent	20,955	24.12	20.4	*2.4 / 3*
	Top 10 Prolific	12,581	14.48	3.4	*3 / 3*
Leurre	Top 10 Persistent	645,169	2.20	300.3	*3.9 / 150*
	Top 10 Prolific	23,530	0.08	2.8	*107/ 150*

The analysis is more interesting for Leurre. We can see for example that the top 10 most persistent attackers exchanged almost 28 times more traffic with the honeypots than the prolific attackers, while at the same time concentrating their efforts on a very small number of honeypots (3.9 honeypot on average for the persistent attackers compared with 107 for prolific ones). This table reconfirms that different attackers are using different strategies when attacking the honeynets.

3.3 The Effect of *Honey* (in the Form of Computing Resources) on the CORP Traffic

So far we have looked at the differences in traffic observed when honeypots are running on different networks. To check the effects that the computing resources (e.g. the deployed applications on the honeypots) have on the traffic we have compared the two honeynet deployments in the CORP network (see Section 2.2 for a description of the setup). Table 4 shows a summary of the results. The main findings are as follows:

- The number of attacking IPs observed in the two networks is very similar, but the total number of packets exchanged by these IPs is 3.5 times higher in the CORP honeynet with applications.
- Most of the packets (82.02% for the honeynet *with* applications and 93.49% for the honeynet *without* applications) are launched by 714 attacking IP addresses which reappear in both honeynets.

Table 4 – Comparison of the two honeynets in the CORP network: *with* and *without* applications in the honeypots.

		Overlap Counts	Single HN counts	**Ratio of totals**
CORP with Applications	Attacking IPs Count	*714*	*1437*	***49.69%***
	Packet Count	282,845	*344,858*	***82.02%***
	Avg. per Attac. IP	***396.14***	***239.98***	***N/A***
CORP no Applications	Attacking IPs Count	*714*	*1442*	***49.51%***
	Packet Count	**92,834**	**99,302**	***93.49%***
	Avg. per Attac. IP	***130.02***	***68.86***	***N/A***

4. Initial Exploratory Models

In this section we give details of two exploratory models that have been developed based on the data presented so far.

4.1 Preferential Attack Model

Using the data for the 10,000 most persistent[2]

[2] Apart from CORP, where there were 1,752 attacking IPs in total. Hence, all CORP attacking IPs were included in the analysis.

attackers we have calculated the distribution of attack frequency and the attack size measured by the amount of traffic from the source (see Fig. 6). Each of the networks shows a power law between the attack probability P per honeypot and size of attack (k). The power law is given by:

$$P(k) \propto k^{-\gamma}$$

with γ between 2.5 and 3.6. The power law (>2) suggests a scale free, preferential attachment mean-field model, such as that of [15] and [16], where the probability of the next attack coming from source k is proportional to amount of traffic seen previously for that attacker (n). A preferential attachment model with a constant of proportionality ($a+n$) would give $\gamma = 3 + a$ with $a > -1$ [16]. It would also predict the constant of proportionality to be proportional to the square of the total amount of attack traffic. It is tempting to interpret the constant a in terms of *background radiation* [2].

The results have implications for the design of adaptive defence strategies.

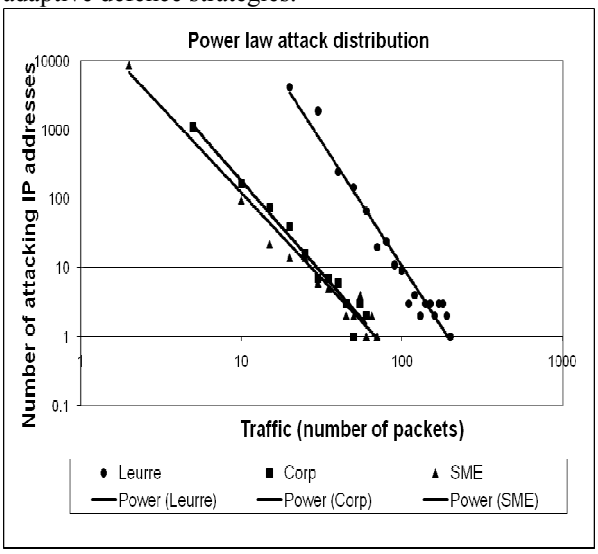

Fig. 6 - Attack frequency vs. Size.

4.2 Empirical Model for the Precursor Attacks

To find out whether attacks on one honeynet could be used as a precursor for an attack on another honeynet we looked at the distribution of time between the earliest Snort IDS (Intrusion Detection System) alerts for either of the two honeynets (SME or CORP) until the first alert on the second honeynet. We did this for the observation period 21/May/2007 to 22/July/2007 for CORP and SME.

Fitting Weibull distribution to the data, we found that square root of this time has approximately an exponential distribution (Fig. 7) with the mean 7.41 for all reappearing attacker IPs on both networks (Kolmogorov-Smirnov test p-value is 0.138).

We also found that the distribution of square root of time from the first packet exchange with either honeynet until the first packet exchange with the remaining honeynet for all reappearing IP addresses is also approximately exponential (Fig. 8) with mean 9.96 (Kolmogorov-Smirnov test p-value is 0.506). Thus, an attack on one honeynet could be seen as a precursor for an attack on the other. This empirical model may therefore help network administrators to predict the time of an attack in their network given knowledge of an attack on another network.

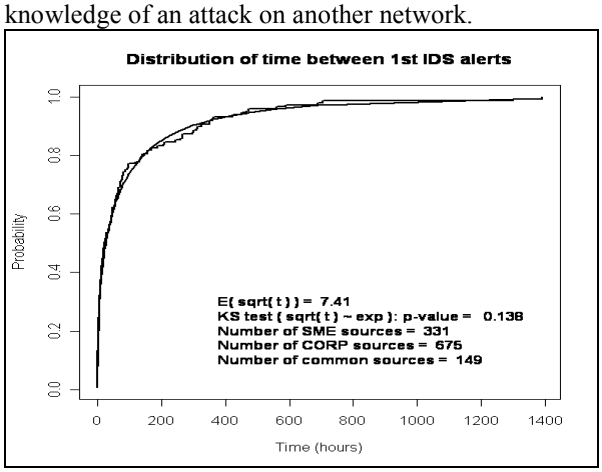

Fig. 7 – Empirical and fitted distributions of absolute values of time difference between 1st IDS alerts from reappearing attacker IPs.

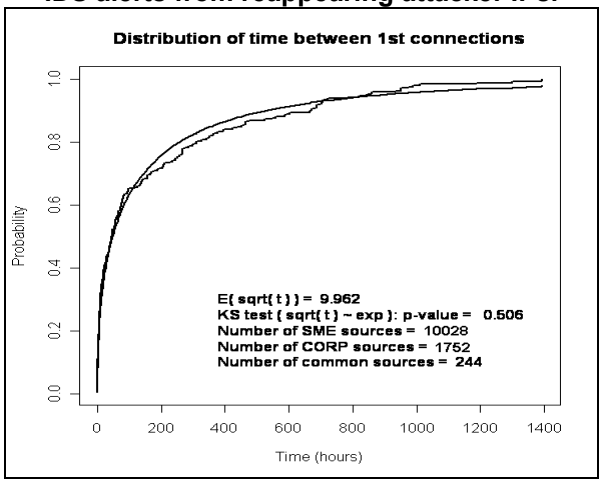

Fig. 8 - Empirical and fitted distributions of absolute value of time difference between 1st packet exchanges from all reappearing attacker IPs.

5. Related work

Numerous publicly available sources publish security related data. Examples include SysAdmin, Audit, Network, Security (SANS) [17], Carnegie

Mellon University's Computer Emergency Response Team (CERT) [18], Honeynet Project [4] and Cooperative Association for Internet Data Analysis (CAIDA) [19]. The last two sources also provide various network and security tools. One of the problems with analyzing data from these sources is that the data is summarized and, thus, it is difficult to perform more specific or detailed analysis if the required data is not available in the summarized reports.

A related honeynet architecture was built by Michel Cukier's team at the University of Maryland (the architecture they developed is outlined in [20]). The main difference between the architecture developed in [20] and that in [4] (which we have used) is in the way the data control and data capture is performed. In [4] all the tools are grouped in a single installation CD making the deployment and configuration much easier. In [20] the data control and data capture is much more distributed. Findings and analysis of the results derived from the honeynet deployment have been detailed in a few papers by Cukier and his team, including: [20] in which it was observed that (using their definitions of port scan, vulnerability scan and attack) port scans should not be considered as precursors to attacks; [21] in which the authors provide empirical and statistical analysis of classifying attacks directed to Windows port 445 (which Short Message Block (SMB) protocol uses), concluding that a criterion as simple as the total number of bytes per connection is very good for separating different attacks on this port, whereas number of packets per connection and connection duration are not so good; or [22] in which the authors analyzed the attacker behavior that follows a successful compromise on Secure Shell (SSH) protocol.

A number of papers have also been published based on the analysis of data from the same Leurre.com source that we have also used in this paper (though these papers use different observation periods than what we report). [14] report initial analysis performed on the data collected in Leurre.com distributed honeypots network. Their results show that 5% of the attack sources are observed on at least two honeypots, most of the attacks are destined to Windows machines and attacks that targeted all three virtual machines on a site were performed in a numerical (increasing or decreasing) order of the respective IP addresses. These findings are also confirmed by the analysis we did on the Leurre.com data for the period reported in this paper, 21/May/2007-22/July/2007 (although we found that when honeypots from the same network are scanned, the scan tends to happen in sequential *ascending* order). [23] report findings collected during a 4 month period using three virtual honeypots deployed with VMware software. Most of the attacks originated from three countries: Australia, Netherlands and USA. The first two are rarely reported as "attackers'" country of origin in computer security reports. The authors speculate that the targets are chosen at random since each one of them is attacked by approximately third of all attacking IPs. The most targeted port is 139, used for Windows NetBIOS protocol. [24] described ScriptGen tool which is used for deployment of medium-interaction honeynets. The tool enables richer communication with attackers than *honeyd* (used currently by the Leurre.com distributed honeypots deployment) without imposing maintenance cost and risk of high interaction honeynets. Initial tests with the SMB protocol are described. [25] report on findings concerning a deployment of a high-interaction honeypot to which only SSH connections were allowed. The main goal of the experiment was to examine the behaviour of the attackers who successfully compromise a machine. They identified that *dictionary attacks* were common. By looking at the pattern of intrusions they have identified both humans and automatic programs as attackers.

The research detailed in [26] used a deployment of a generation I honeynet [27] to monitor traffic and identify malicious activities in a corporate environment provided by The Georgia Institute of Technology. The reported monitoring period lasted for six months. In this period 16 machines, which were outside of the honeynet's IP address space, have been discovered as compromised. Beside likely worm propagation attacks, the honeynet helped in the identification of a system with a compromised password. The authors propose the use of a distributed honeynet, controlled by a network separate from the corporate one, as a means for further security enhancement.

Similarly to the exploratory models in Section 4, the work of [28] presents some preliminary analyses and modelling techniques to better understand malicious activities on the Internet and the corresponding attacker strategies. The authors are concerned with time-evolution modelling of number of attacks as well as potential correlations among attacks on geographically dispersed platforms. They investigated the distributions of attacks on a single platform and, also, the propagation of attacks through different platforms. The analyses are based on the data collected in Leurre.com. The most interesting results are as follows:

- Trends at a local level do not necessarily follow the global trend.
- Attack processes differ among platforms and they are governed by specific factors.
- Heavy-tailed power law distribution characterises the number of attacker IPs as a function of number of attacks per platform – few attacker IPs are responsible for majority of attacks.

- The observed times between attacks on a particular platform are best characterised using a mixture model combining a Pareto and an exponential distribution (NB: the exploratory model in Section 4.2. is concerned with inter-platform attacks).

6. Conclusions

In this paper we have reported empirical observations and exploratory data analysis based on data from three different networks: two honeynets running in a corporate and an SME environment (both deployed in the UK) and a distributed honeypots network of 150 honeypots running in 50 sites around the world. We observed that both the total number of packets launched by attackers and the total number of attackers per honeypot differed in the three networks: Leurre.com distributed honeypots network had the highest average number of packets per honeypot and by far the highest number of distinct attacker IPs. We compared the traffic from IP addresses that attack more than one network. We saw that over 60% of attacking IPs observed in the corporate honeynet (which contributed more than 90% of the traffic in this honeynet) were also observed in the Leurre.com network. This percentage is smaller when we compare the SME and Leurre.com (34.2% of SME attacking IPs were observed in Leurre.com), CORP with SME or all three.

There is some consistency in the top ranking countries of origin of the attackers in the three networks: IP addresses from the United States are the most frequent and exchange the most packets with the honeypots in any of the three networks, with Chinese IP addresses also being in the top 5.

We analyzed the prolificacy of the attacking IP addresses with respect to the number of honeypots they attack in the Leurre.com network. We observed that among the top 10 most prolific attacking IPs 7 are from China. This would indicate that Chinese IP addresses are more prone to scanning a large number of hosts rather than concentrating their attacks against single hosts. When we look at the times in which the Leurre.com honeypots are discovered (i.e. the times between first packet exchanges for a given attacking host and any of the honeypots in Leurre.com network) we observed seemingly different strategies being employed by the attackers: most of the attackers seem to be going through high number of hosts in Leurre.com very quickly before the rate of discovery of new hosts drops, whereas for some the opposite is true (they discover very few new hosts initially and then the rate increases). We also observed that the total number of packets exchanged with the honeypots in Leurre.com is relatively low for the prolific hosts. This may suggest that these hosts are involved in "intelligence gathering" at this stage: probing with very few packets which hosts are alive and what services they are running before they decide to launch more concentrated attacks against a few hosts they determine to be more vulnerable.

A few other interesting observations include:

- the persistence of attackers (i.e. the number of packets that an attacker will launch against a network) seems to follow the Power law: a proportionally small number of attackers for each given network are responsible for very large amount of the malicious traffic.
- a simple preferential attachment model was given, which predicts that the probability of the next attack coming from an attacking IP is proportional to the number of attackers.
- initial analysis with part of the data for the corporate and SME honeynets suggests that square root of time for an attacker who attacks one of these honeynets to attack the other is exponentially distributed with mean of approximately 10.
- the scanning sequence of the hosts in the same network seems to be predominantly in sequential ascending order. This phenomenon was also reported in a related study of the Leurre.com network data [14] (although the ordering observed in [14] was either ascending or descending).

We are working to extend the preliminary analysis reported here to develop statistical models for attack behaviour e.g. the propagation time of an attack from one network to another. We are interested in experimenting with adaptive defence strategies for a single network based on measuring traffic to itself and some statistical models for the internet as a whole. We are also interested in inferring global behaviour from a comparative analysis of these different networks. Other possibilities for future work include:

- repeating the analysis for a different period with these networks and comparing the results. It may especially be interesting to analyze what happens with the IP addresses that were identified as "persistent" and "prolific" in this period. Will the "prolific" IP addresses (i.e. the ones that were scanning a lot of hosts) launch more "persistent" attacks against a smaller number of hosts?
- define precise null hypotheses that can be tested with the observations.
- define initial exploratory models and more refined explanatory probabilistic models for predicting variables of interest, e.g. the time between attacks in different networks.

Acknowledgments

This work has been supported by a Strategic Development Fund (SDF) grant from City University London. We thank Bev Littlewood for reviewing an earlier version of the paper. We would also like to thank the City of London Police for their advice and support on forensic and risk assessment issues.

References

1. Leurrecom, *"Leurrecom Honeypot project",* 2007; Available from: http://www.leurrecom.org/.
2. Ruoming, P., Y. Vinod, et al., *"Characteristics of internet background radiation",* in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. 2004, ACM: Taormina, Sicily, Italy.
3. Bloomfield, R., I. Gashi, et al., *"Building and Deploying a Honeynet Infrastructure in a Corporate Network"*. 2007.
4. Honeynet-Project-Research-Alliance, *"Honeynet Project",* 2007; Available from: http://www.honeynet.org/.
5. Honeynet-Project-Research-Alliance, *"Know Your Enemy: Honeywall CDROM Roo",* 2007; Available from: http://www.honeynet.org/papers/cdrom/roo/index.htm.
6. Software-Foundation-Apache, *"Apache HTTP Server",* 2007; Available from: http://www.apache.org/.
7. PostgreSQL, *"PostgreSQL",* 2007; Available from: http://www.postgresql.org.
8. OpenOffice.org, *"OpenOffice",* 2007; Available from: http://www.openoffice.org/.
9. Mozilla, *"Thunderbird Mail client",* 2007; Available from: http://www.mozilla.com/en-US/thunderbird/.
10. TPC, *"TPC Benchmark C, Standard Specification, Version 5.0."* 2002; Available from: http://www.tpc.org/tpcc/.
11. GetSafeOnline, *"Get Safe Online - Expert Advice for everyone",* 2007; Available from: http://www.getsafeonline.org/.
12. Monkey.Org, *"Honeyd",* 2007; Available from: http://www.honeyd.org/.
13. Gashi, I. and V. Stankovic, *"Empirical results of comparing traffic from two honeynets and a distributed honeypots project ",* 2008; Available from: http://www.csr.city.ac.uk/people/ilir.gashi/Security/Whitepapers/.
14. Pouget, F., M. Dacier, and V.H. Pham, *"Leurre.com: on the advantages of deploying a large scale distributed honeypot platform",* in *E-Crime and Computer Conference (ECCE'05)* 2005. Monaco.
15. Barabasi, A.L. and R. Albert, *"Emergence of Scaling in Random Networks",* Science, 1999. 286: p. 509-512.
16. Durrett, R., *"Random Graph Dynamics"*. 2007: Cambridge University Press.
17. SANS, *"SysAdmin, Audit, Network, Security",* 2007; Available from: http://www.sans.org/.
18. CERT, *"Carnegie Mellon University's Computer Emergency Response Team",* 2007; Available from: http://www.cert.org/.
19. CAIDA, *"Cooperative Association for Internet Data Analysis",* 2007; Available from: http://www.caida.org.
20. Panjwani, S., S. Tan, et al., *"An Experimental Evaluation to Determine if Port Scans are Precursors to an Attack",* in *Dependable Systems and Networks (DSN-05)*. 2005. Yokohama, Japan: IEEE Computer Society Press, p. 602-611.
21. Cukier, M., R. Berthier, et al., *"A Statistical Analysis of Attack Data to Separate Attacks",* in *IEEE Dependable Systems and Networks (DSN'06)*. 2006. Philadelphia, Pennsylvania, USA: IEEE Computer Society, p. 383-392.
22. Ramsbrock, D., R. Berthier, and M. Cukier, *"Profiling Attacker Behavior Following SSH Compromises",* in *IEEE Dependable Systems and Network (DSN'07)*. 2007. Edinburgh, UK: IEEE Computer Society, p. 119-124.
23. Pouget, F., M. Dacier, and H. Debar, *"Honeypots, a practical mean to validate malicious fault assumptions",* in *International Symposium Pacific Rim Dependable Computing Conference (PRDC'04)*. 2004. Tahiti, French Polynesia.
24. Leita, C., K. Mermoud, and M. Dacier, *"ScriptGen: an automated script generation tool for honeyd",* in *21st Annual Computer Security Applications Conference (ACSA'05)*. 2005. Tucson, USA.
25. Alata, E., V. Nicomette, et al., *"Lessons learned from the deployment of a high-interaction honeypot ",* in *European Dependable Computing Conference (EDCC'06)*. 2006. Coimbra, Portugal: IEEE Computer Society, p. 39-46
26. Levine, J., R. LaBella, et al., *"The Use of Honeynets to Detect Exploited Systems Across Large Enterprise Networks",* in *IEEE Workshop on Information Assurance*. 2003. United States Military Academy, West Point, NY, USA.
27. Honeynet-Project-Research-Alliance, *"Know Your Enemy: Honeynets",* 2005; Available from: http://www.honeynet.org/papers/honeynet/index.html.
28. Kaaniche, M., Y. Deswarte, et al., *"Empirical analysis and statistical modeling of attack processes based on honeypots",* in *IEEE Dependable Systems and Networks (DSN'06)*. 2006. Philadelphia, Pennsylvania, USA: IEEE Computer Society, p. 119-124.